DOCUMENT RESUME

ED 038 679 CG 005 225

AUTHOR	Cox, Richard C.
TITLE	Evaluative Aspects of Criterion-Referenced Measures.
INSTITUTION	American Educational Research Association, Washington, D.C.; Pittsburgh Univ., Pa.
PUB DATE	2 Mar 70
NOTE	8p.; Paper presented at American Educational Research Association Convention, Minneapolis, Minnesota, March 2-6, 1970
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.50
DESCRIPTORS	Changing Attitudes, *Criteria, *Critical Incidents Method, *Evaluation Methods, Measurement, *Measurement Techniques, Teaching Techniques, *Test Construction, Testing, Tests

ABSTRACT

The plea has been made to interpret the concept of criterion-referenced measurement in a broader sense, so that the idea will be utilized in more ways in educational measurement. While applications of the concept have been suggested mainly for programs of individualization, there is no reason to limit the ideas of criterion-referenced measurement to such application. There needs to be some thought given to how the criterion-referenced concept can be applied to the typical teacher-made test as well as to standardized tests. Also, if the idea is to be accepted, some alternative to the traditional approaches to reliability, validity, and item analysis procedures must be investigated. (KJ)

Evaluative Aspects of Criterion-Referenced Measures*

Richard C. Cox
University of Pittsburgh

While suggestions for criterion-referenced measurement have been available for some time it has been only in the past few years that the notion has been pursued with interest. Unfortunately the suggested usages reflect only a quite narrow definition of criterion-referenced measurement, when actually there are many ways in which the concept can provide meaningful test results. What is needed is a broader definition of the term which will be accepted and used by more people in the educational measurement field. If this acceptance is to be forthcoming there needs to be a reexamination of some of the evaluative aspects of measurement, such as reliability, validity, and item analysis procedures. Traditional approaches to these concepts may not be easily adaptable to criterion-referenced measurement; therefore, suggestions for gathering evaluative data should be explored.

The major distinction between norm-referenced and criterion-referenced measurement concerns the type of information provided. If, for example, an achievement test is administered in order to provide information about the performance of a pupil compared with that of other pupils, the measurement is said to be norm-referenced. Pupils are ordered with respect to each other or to some well defined norm group. While it is also possible to order pupils using criterion-referenced measurement, the more valuable information to result in this case is each pupil's performance

*Prepared for AERA Symposium, "Criterion-Referenced Measurement: Emerging Issues", Minneapolis, Minnesota, March 1970.

relative to some specific standard. The essence of criterion-referenced measurement is in the specificity of information yielded in terms of pupil performance relative to some criterion. The following example illustrates this distinction between norm and criterion-referenced measurement.

A graduate measurement examination is constructed to test the concepts of reliability, validity, and item analysis. After the test is administered the scores can be interpreted in either a norm or a criterion-referenced sense. If the instructor wishes to assign grades to individuals on the basis of the results, he may use the class as its own norm group and assign A's to those pupils with scores at least one standard deviation above the mean of the group, C's to those scoring at least one standard below the mean, with B's for the remaining scores. The grading could also be based upon certain percentile points, stanine scores, or other standard scores. Other appropriate norm groups may also be used, for example, the previous class. What is important in the measurement procedure is that the performance of individuals is compared with reference to some norm.

On the other hand, the test may be interpreted in a criterion-referenced manner. It may be the case that the instructor has established a criterion level of performance (such as 80 percent of all items answered correctly within each area) as an acceptable minimum standard. Pupils achieving this criterion performance are considered ready to study further measurement topics. Pupils who have not achieved the minimum standard may be given some form of remedial work or may just be considered as not having met the requirements. The important point here is that some specified performance level is the criterion and that the instructor is able to determine just what each pupil does or does not know in

reference to each topic tested.

This example illustrates another often neglected point. It is possible for a single test to yield both norm-referenced and criterion-referenced information. (In fact, it has been suggested in an earlier paper that with certain adaptations, criterion-referenced data can be obtained from a typical standardized achievement test which has been designed to yield only norm-referenced information.) This is illustrative of the notion that criterion-referenced measurement need not be operationally defined in such a restricted sense.

Most available examples of criterion-referencing have been associated with individualized instructional programs. Coulson and Cogswell (1965) discuss the need for criterion-referencing in regard to the use of programmed materials utilized in individualized instructional systems. Glaser and Cox (1968) also suggest the use of criterion-referenced measures in individualized instruction models where the evaluation instruments must differentiate between groups of pupils who have mastered certain units of instruction and those who have not. While much of the impetus behind the discussion of criterion-referenced measurement has come from innovative instructional programs requiring precise specifications of instructional objectives, it would be unfortunate if this restricts further exploration of the concept.

Another restriction, recently discussed by Popham and Husek (1969), is that some of the traditional approaches to measurement theory are either not applicable or are difficult to adapt to the criterion-referenced framework. If the idea of criterion-referencing is to be accepted and used in educational measurement, there needs to be discussion and suggestions for some alternatives to reliability, validity, and item analysis procedures. The following discussion

presents some suggestions for ome new approaches to these concepts.

When an achievement test is constructed as a norm-referenced measure the test items are written or selected to maximize differences between individuals. Maximum discrimination is desirable to obtain the variability necessary for ranking individuals. Similarly, most of the empirical methods for reporting reliability or validity indices require that the range of scores not be greatly restricted. Those test items which are answered correctly by all examinees or by no examinee will have difficulty levels of 1.00 and 0.00 respectively, and cannot discriminate in the usual sense between those scoring high and low on the total test. Items with characteristics like these will not contribute to test variance and will therefore be eliminated from a test designed to discriminate between individuals since their only effect is to add or subtract, a constant value to or from every score.

A study by Cox and Vargas provides some data relevant to this point. Two discrimination indices were computed for items on tests which had been administered both as pre and post tests. The question of interest was the extent to which the two methods of item analysis yield the same relative evaluation of items. One index was computed using the common upper minus lower groups technique, thus providing information on how well each item discriminated between these groups. The second index involved both the pre and post test and was computed by subtracting the percentage of pupils who passed the item on the pre-test, from the percentage who passed the item on the post test. This index provided discrimination information between pre and post test groups, indicating items useful for pre-test diagnosis. Results of the comparison between the two indices indicated that some items which are highly desirable for the pre-post test discrimination would be discarded by the typical

item selection techniques, because they fail to discriminate among individuals taking the test. It was concluded that the pre and post test method of the item analysis produced results sufficiently different from traditional methods to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures.

While this study proposed an approach to item analysis for criterion-referenced measures, it has application only in the pre-post test situation. There is certainly a need for some development work on item analysis procedures when only one test administration is possible. (Perhaps we will hear about such work later in this symposium.) Consideration like these must be given attention if the criterion-referenced concept is to be broadened.

Exactly what happens to the concepts of validity and reliability when applied to criterion-referenced measures is not clear. It may be the case that the usual coefficients are adequate; it is also quite possible that the criterion-referenced measure will not yield sufficient variability to make these typical coefficients meaningful. Consider, for example, the case where all individuals taking the test answer all of the items correctly. Not only will the items have no discriminating power among individuals, but also due to the absense of variance it may not be possible to interpret traditional reliability or validity coefficients.

A study by Cox and Graham (1966) illustrates one way in which reliability may be viewed, given a special type of criterion-referenced measure. They described the development of a sequentially scaled achievement test designed for use in an instructional system within which certain performance objectives can be identified as being sequential in nature. Theoretically, in this situation it seems possible to construct a test in such a way that the pupil

answers all items up to a certain point (his level of attainment) and misses all items beyond that point. The test would be scaled in the Guttman sense, the total test score indicating the response pattern of the individual. (This is a good example of criterion-referenced measurement where the test information is quite specific with reference to examinee performance according to some criteria.) The analysis of a group of such scores yeilds a coefficient of reproducibility which indicates how well an individual's response pattern can be reproduced from a knowledge of his total score. This coefficient, while usually considered as a verification of the arrangement of items, might also be used as a type of reliability estimate across all individuals taking the test. The pitfalls of using reproducibility as a reliability estimate for achievement tests has not yet been explored.

Validity is a major concern in criterion-referenced measurement. The emphasis on obtaining information specific to pupil performance with reference to some criterion makes obvious the need for validity. As always, these validity estimates must be determined by the purpose for which the test is being used. Certain uses will dictate certain necessary validities. In general, however, criterion-referencing itself suggests that validity must depend upon the correspondence of the test items with the objectives to which the test is referenced. Criterion-referenced tests then, must provide information in terms of specific behavior. Thus the test items must be constructed for, or matched to, goals of instruction. The desired measurement must provide information in terms of pupil performance relative to some criterion and therefore demands a rigorous validation procedure.

The use of experimental techniques to establish the validity of a criterion-referenced measure should be investigated (i.e. construct-validation procedures). An example might be that if teaching

techniques have been effective, then a pupil who has worked through a given unit of content should attain a higher score on a post test than a pupil who has not yet been exposed to the unit content. Many such operational definitions could be examined under the heading of construct validity.

In summary, the plea has been made to interpret the concept of criterion-referenced measurement in a broader sense, so that the idea will be utilized in more ways in educational measurement. While applications of the concept have been suggested mainly for programs of individualization, there is no reason to limit the ideas of criterion-referenced measurement to such application. There needs to be some thought given to how the criterion-referenced concept can be applied to the typical teacher-made test as well as to standardized tests. Also, if the idea is to be accepted, some alternative to the traditional approaches to reliability, validity, and item analysis procedures must be investigated.

REFERENCES

Coulson, J.E. and Cogswell, J.F. "Effects of individualized instruction on testing," Journal of Educational Measurement, 2 (1), 59-64, 1965.

Cox, R.C. and Graham, G.T. "The development of a sequentially scaled achievement test." Journal of Educational Measurement, 3 (2), 147-150, 1966.

Cox, R.C. and Vargas, J.S. "A comparison of item selection techniques for norm-referenced and criterion-referenced tests." Paper read at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, February 1966.

Glaser, R. and Cox, R. C. "Criterion-referenced testing for the measurement of educational outcomes." In R.A. Weisgerber (editor), Instructional process and media innovation. Chicago, Illinois: Rand McNally and Co., 1968, pp. 545-550.

Popham, W. J. and Husek, T.R. "Implications of criterion-referenced measurement," Journal of Educational Measurement, 6, (1), 1-9, 1969.